Statistical Downscaling of Rare Events. Application to Storm Forecast

L.M. Oviedo (1), M.R. Pons (1), D. San-Martín (2), R. Ancell (1)
(1) Agencia Estatal de Meteorología (AEMET), Santander, Spain
(2) Dept. of Applied Mathematics. University of Cantabria

General overview. Motivation

- The aim of this work is to analyse the feasibility of an operative thunderstorm forecast all over Spain using Statistical Downscaling techniques
- Logistic Regression (LR) has been frequently used for single-site storm forecast, using air sounding data (<u>very good quality data!</u>). That's the reason why in published results, skill scores obtained with LR are very high (Monzato,2007, Sanchez et al,2007)
  - BUT, it is very expensive/unoperative to predict all over Spain with sounding data.
- Different Analog methods have been used successfully to predict other multi-site meterological non-rare events, …

Is it skillful to predict thunderstorms as well?

Is it able to outperform LR?

Work scheme

0

ACM DATA: ERA-40 re-analysis
OBS. DATA: 22 Stations from AEMET Network in Northern Spain
(BINARY/daily data)

LOGISTIC REGRESSION

1

We look for a BENCHMARK!

Logistic Regression method

ANALOGS

2

We try to beat the benchmark with analogs.

Analogs system with N PC's and M analogs.

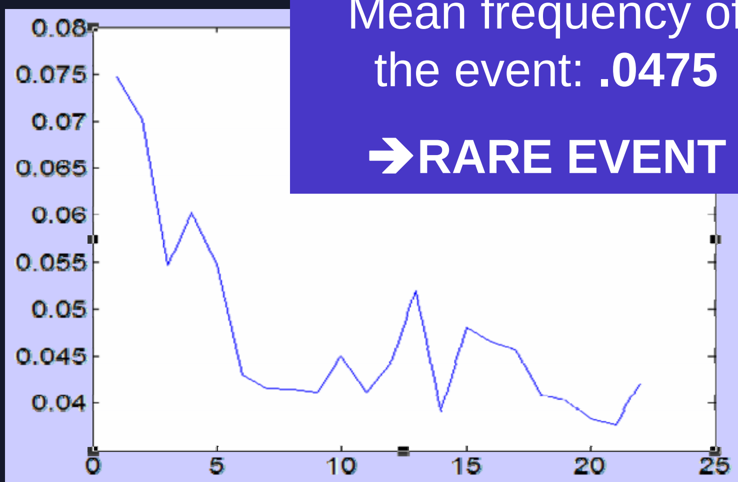
3

- Analogs-logistic comparison. Probabilistic and deterministic validation:
    - ROC curve, Reliability and Resolution
    - HIR,POFD,ORSS,EDSS

0

Data availability

ACM DATA: ERA-40 re-analysis
OBS. DATA: 22 Stations from AEMET Network in Northern Spain (BINARY/daily data)

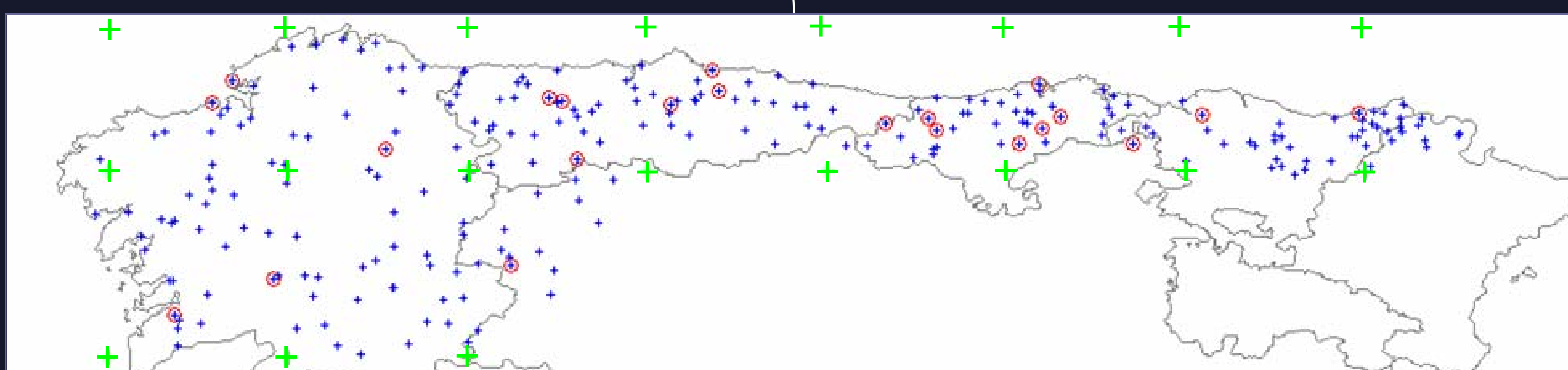


Mean frequency of the event: **.0475**

→RARE EVENT

- Data
  - ACM DATA: ERA-40 re-analysis
  - OBS. DATA: 22 Stations from AEMET Network in Northern Spain (BINARY/daily data)
- Pattern:
  - Variables
    - T,Z,R,U,V at 1000,925,850,775,700,500 hPa at 0,12,24 Z
    - Potential Vorticity at 300hPa, Relative Voriticity at 300 and 500 hPa at 0,12,24 Z
    - Total Column Water at 0,12,24 Z
    - Dew-Point Depresion Index* = T-Td
    - Totals Total* = T850 -2*T500+Td850
    - K Index* = T850 –T500 + DD700

19 nodes*65 fields = 1235 fields.



Work scheme

0

ACM DATA: ERA-40 re-analysis
OBS. DATA: 22 Stations from AEMET Network in Northern Spain (BINARY/daily data)

LOGISTIC REGRESSION

1

We look for a BENCHMARK!

Logistic Regression method

ANALOGS

2

We try to beat the benchmark with analogs.

Analogs system with N PC's and M analogs.

3

•Analogs-logistic comparison. Probabilistic and deterministic validation:

- •ROC curve, Reliability and Resolution
- •HIR,POFD,ORSS,EDSS

LOGISTIC REGRESSION

1

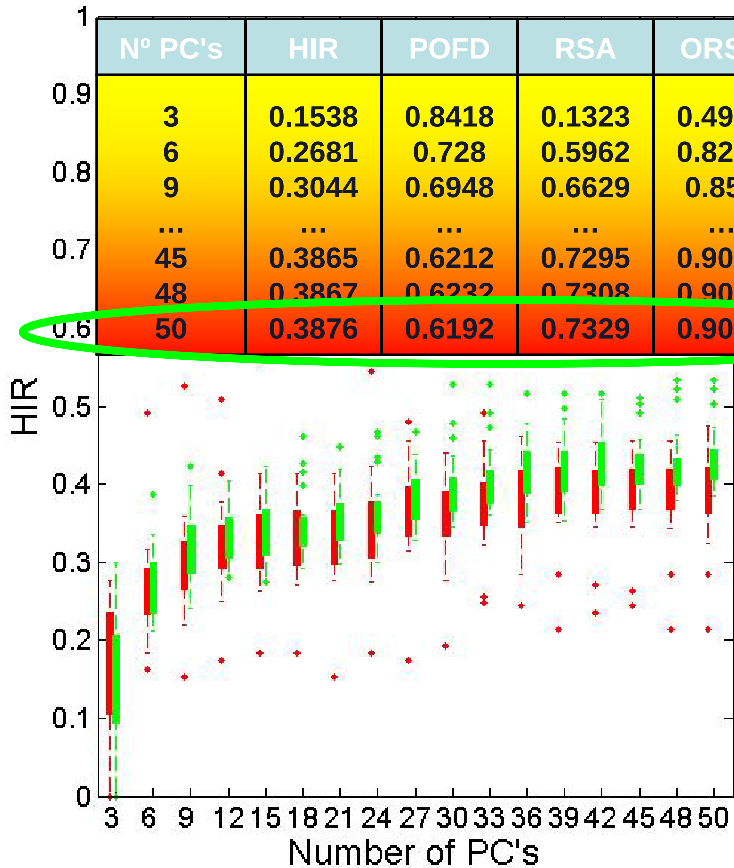
We look for a BENCHMARK!

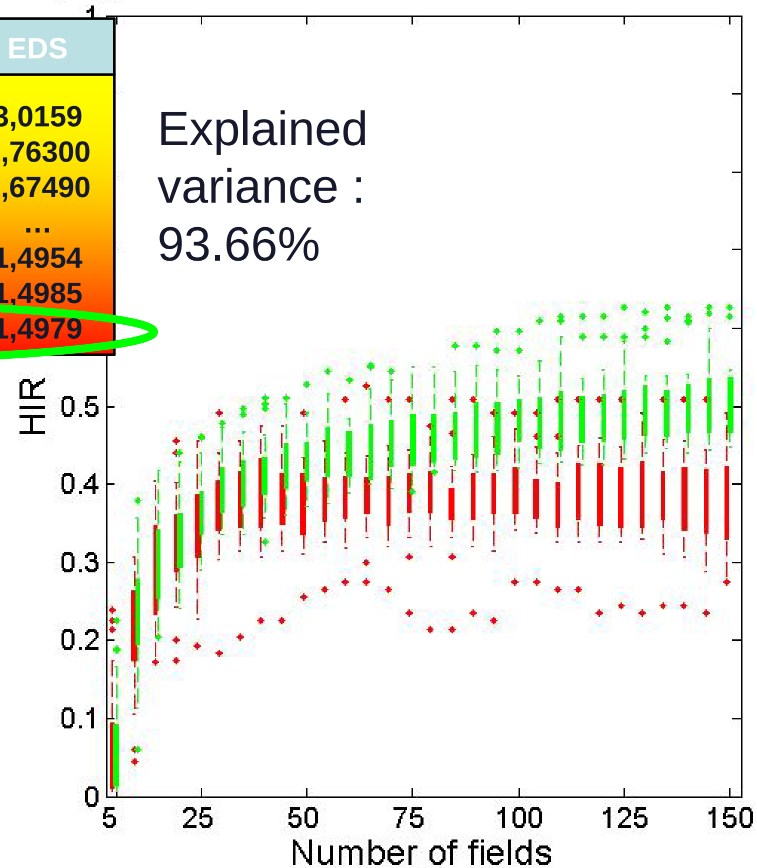
Logistic Regression with N PC's normalized

- Depending on the predictors' input:
  - PC's or ERA-40 Pattern Fields
  - Number of variables.
- Overfitting control
  - Pre-processing techniques (standarize /rescale /normalize) and limiting the predictors' number in order to minimize overfitting.

Test skill vs Train skill in all the 22 stations, varying the number of PC's and number of fields

Nº PC's	HIR	POFD	RSA	ORSS	EDS
3	0.1538	0.8418	0.1323	0.4982	3,0159
6	0.2681	0.728	0.5962	0.8203	1,76300
9	0.3044	0.6948	0.6629	0.855	1,67490
…	…	…	…	…	…
45	0.3865	0.6212	0.7295	0.9067	1,4954
48	0.3867	0.6232	0.7308	0.9063	1,4985
50	0.3876	0.6192	0.7329	0.9077	1,4979





Explained variance : 93.66%

Work scheme

0

ACM DATA: ERA-40 re-analysis
OBS. DATA: 22 Stations from AEMET Network in Northern Spain
(BINARY/daily data)

LOGISTIC REGRESSION

1

We look for a BENCHMARK!

Logistic Regression with 50 PC's normalized

ANALOGS

2

We try to beat the benchmark with analogs.

Analogs system with N PC's and M analogs.

3

- Analogs-logistic comparison. Probabilistic and deterministic validation:
  - ROC curve, Reliability and Resolution
  - HIR,POFD,ORSS,EDSS

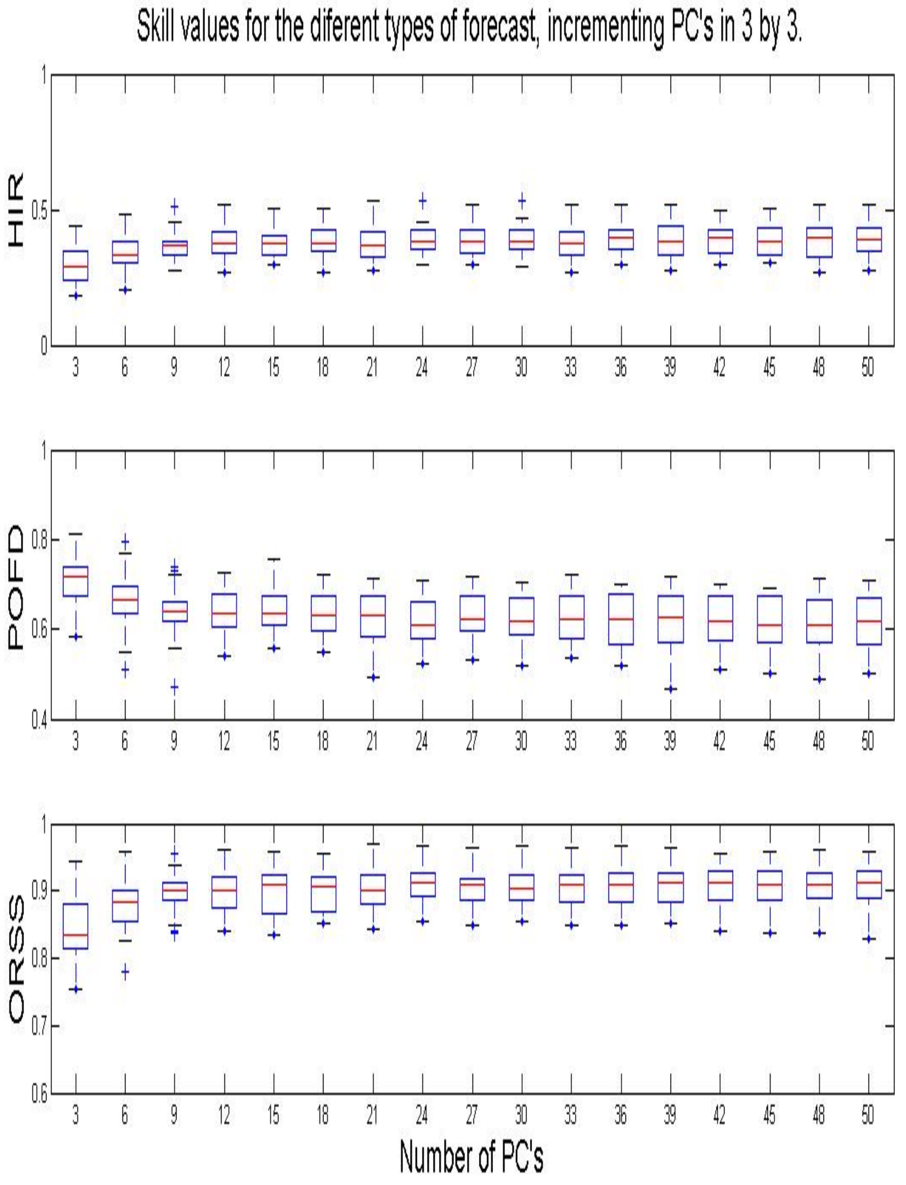
ANALOGS

2

We try to beat the benchmark with analogs.
Analogs system with <u>N</u> PC's and <u>M</u> analogs.

- Depending on the number of PC's: **<u>50 cp's→N=50</u>**
- Depending on the number of analogs: **<u>50 neighbours→M=50</u>**
- Depending on the estimate function: **<u>weighted mean</u>** (weights =>inverse of distance)

Nº neig.	HIR	POFD	RSA	ORSS	EDS
3	0.2927	0.7058	0.4484	0.8609	1,7041
6	0.3396	0.6598	0.5797	0.8934	1,5892
9	0.3652	0.6384	0.6464	0.9043	1,5282
…	…	…	…	…	…
45	0.3896	0.617	0.7863	0.9161	1,4906
48	0.3897	0.6157	0.7887	0.9165	1,4931
50	0.3908	0.6164	0.7882	0.9163	1,4893



Work scheme

0

ACM DATA: ERA-40 re-analysis
OBS. DATA: 22 Stations from AEMET Network in Northern Spain (BINARY/daily data)

LOGISTIC REGRESSION

1

We look for a BENCHMARK!

Logistic Regression with N PC's normalized

ANALOGS

2

We try to beat the benchmark with analogs.

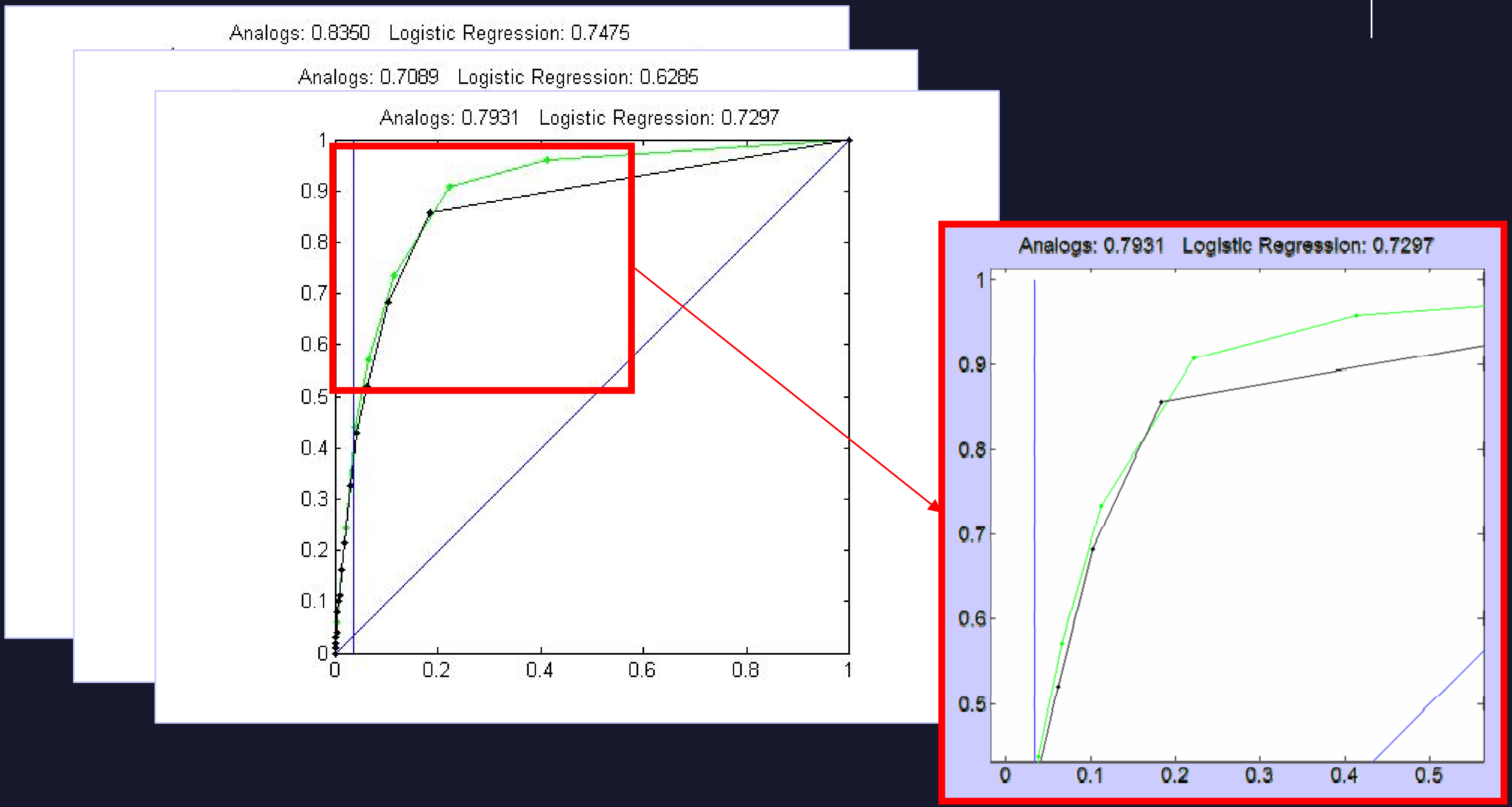
Analogs system with N PC's and M analogs.

3

- Analogs-logistic comparison. Probabilistic and deterministic validation:
  - ROC curve, Reliability and Resolution
  - HIR,POFD,ORSS,EDSS

ANALOGS VS LOGISTIC REGRESSION

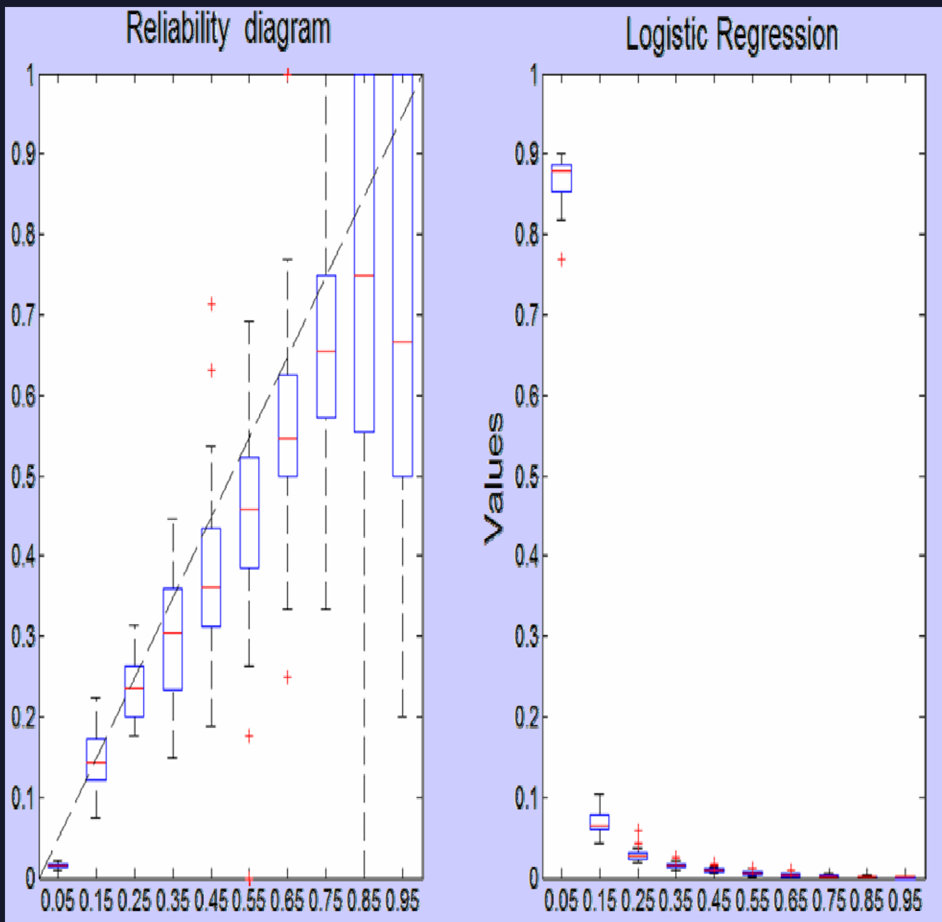
3.1 Probabilistic: ROC Curve

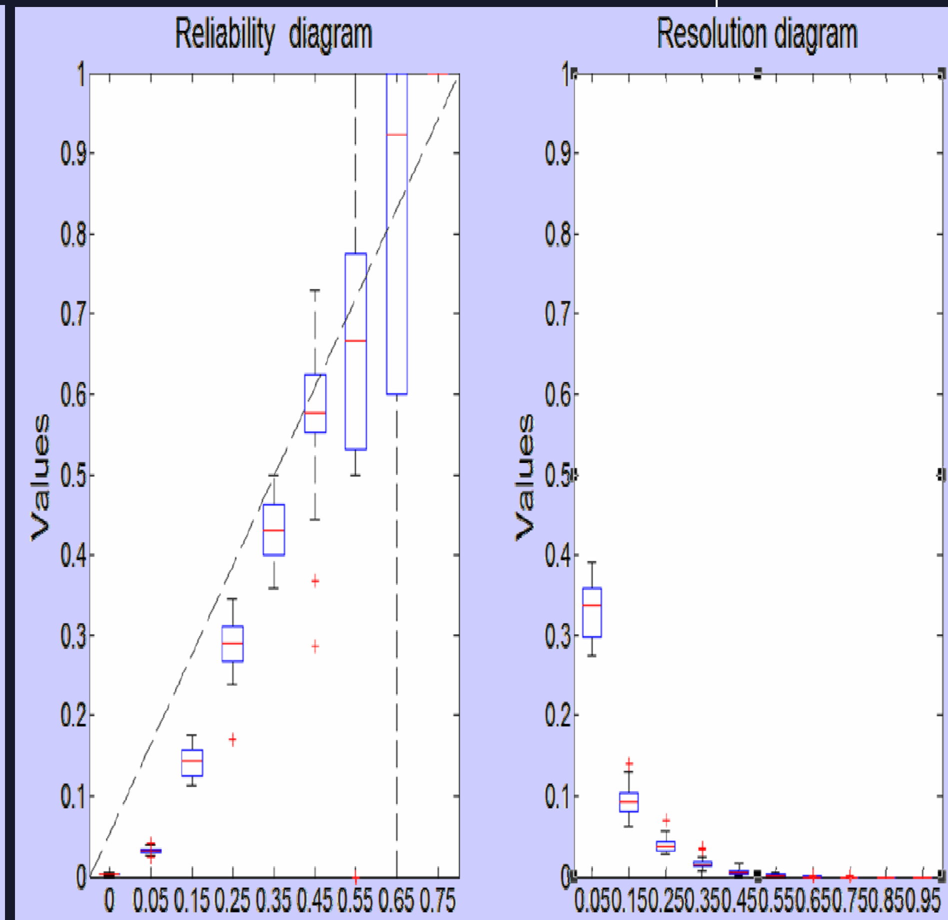


ANALOGS VS LOGISTIC REGRESSION

3.1

Probabilistic: Reliability



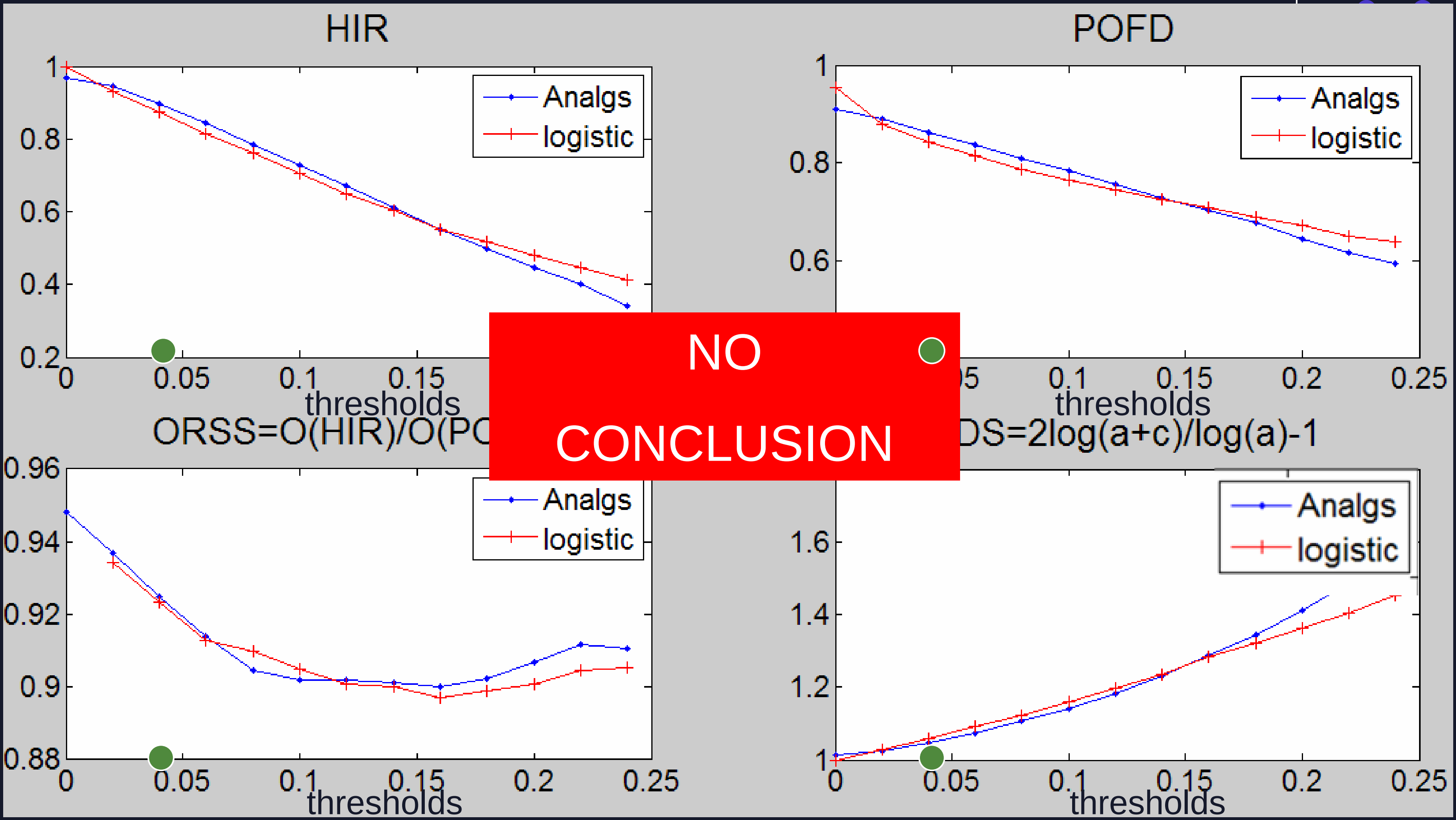


Analogs seems to be slightlier overconfident than RL. Moreover, it detects much better when there is no storm

ANALOGS VS LOGISTIC REGRESSION

3.2

Binary



0

ACM DATA: ERA-40 re-analysis
OBS. DATA: 22 Stations from AEMET Network in Northern Spain
(BINARY/daily data)

LOGISTIC REGRESSION

1

We look for a BENCHMARK!

Logistic Regression with 50 PC's normalized

ANALOGS

2

We try to beat the benchmark with analogs.

Analogs system with 50 PC's and 50 analogs.

3

•Analogs-logistic comparison.
Probabilistic and deterministic validation:

- •ROC curve and Reliability
- •HIR,POFD,ORSS,EDSS

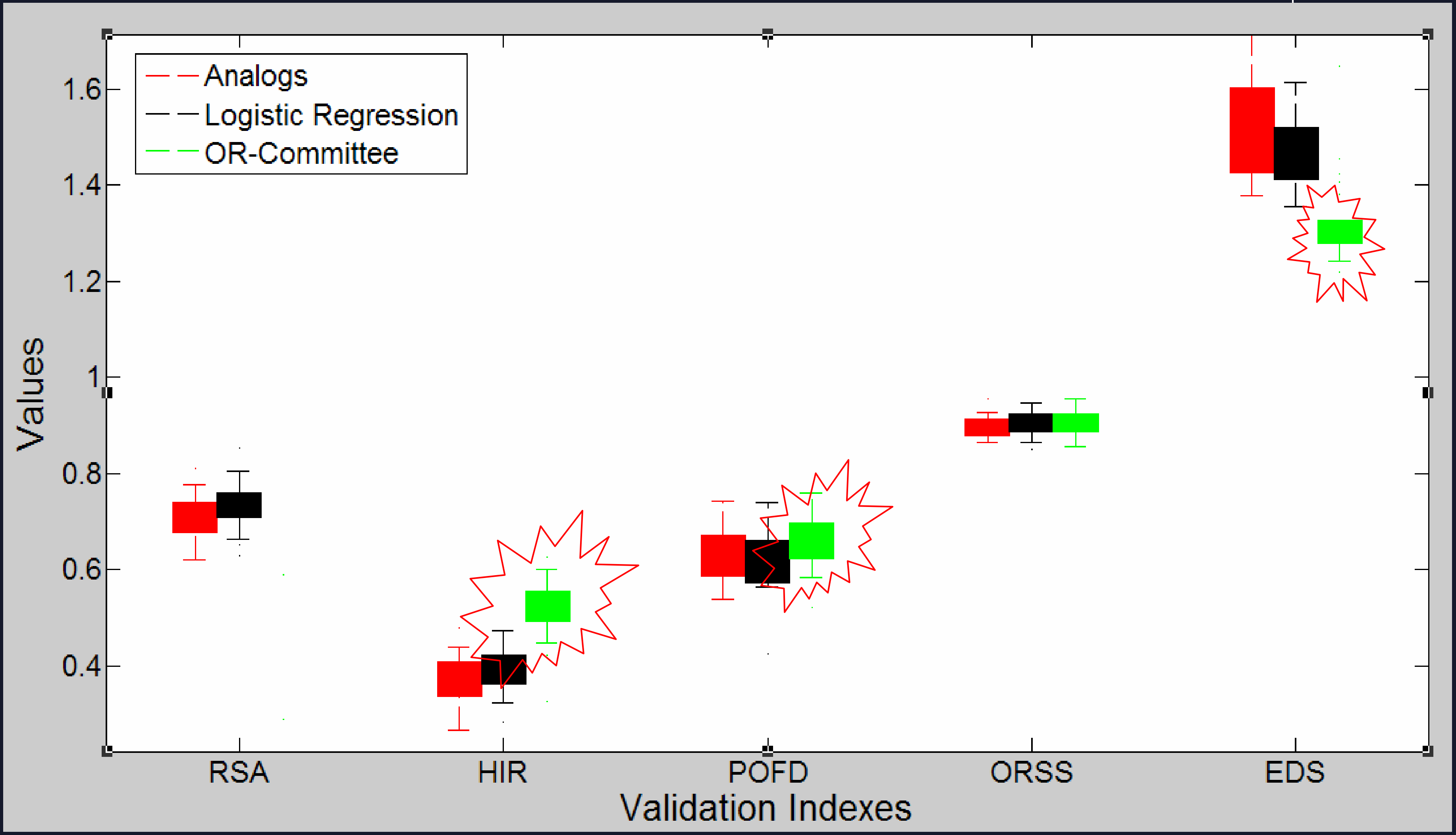
4

¿What if we mix both methods(Expert Committee)?

EXPERTS COMMITTEE

4

Summary



Comparison Logistic Regression vs Analogs method

- Conclusion.Why should we use analogs??
  - There is no clear conclusion about which system is better in an objective way, but we know that logistic regression has few parameters; the method is defined once you set the predictors. Analogs procedure can be modified changing the analogs number, number of neighbours, estimation function(mean, weighted mean, percentile…)
  - Once you chose the predictors set, logistic regression needs to use different coefficients for each station. Analog methods need only one configuration for all the network, so, it's easier to implement analogs in an operative forecast with such a large network of stations.
  - The OR-Committee has resulted the best method. Now we have a lot to do, working on:
    - Analogs technique improvement modifiying the Train period to make the event not rare.
    - More complex "Experts committees" including Bayesian networks or/and Neural networks.

Thank you for your attention!!